View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Research Repository UCD

EXPLORING CUSTOMER REVIEWS FOR MUSIC GENRE CLASSIFICATION AND EVOLUTIONARY STUDIES

Sergio Oramas[1], Luis Espinosa-Anke[2], Aonghus Lawlor[3], Xavier Serra[1], Horacio Saggion[2]
[1]Music Technology Group, Universitat Pompeu Fabra
[2]TALN Group, Universitat Pompeu Fabra
[3]Insight Centre for Data Analytics, University College of Dublin
{sergio.oramas, luis.espinosa, xavier.serra, horacio.saggion}@upf.edu, aonghus.lawlor@insight-centre.org

ABSTRACT

In this paper, we explore a large multimodal dataset of about 65k albums constructed from a combination of Amazon customer reviews, MusicBrainz metadata and AcousticBrainz audio descriptors. Review texts are further enriched with named entity disambiguation along with polarity information derived from an aspect-based sentiment analysis framework. This dataset constitutes the cornerstone of two main contributions: First, we perform experiments on music genre classification, exploring a variety of feature types, including semantic, sentimental and acoustic features. These experiments show that modeling semantic information contributes to outperforming strong bag-of-words baselines. Second, we provide a diachronic study of the criticism of music genres via a quantitative analysis of the polarity associated to musical aspects over time. Our analysis hints at a potential correlation between key cultural and geopolitical events and the language and evolving sentiments found in music reviews.

1. INTRODUCTION

With the democratisation of Internet access, vast amounts of information are generated and stored in online sources, and thus there is great interest in developing techniques for processing this information effectively [27]. The Music Information Retrieval (MIR) community is sensible to this reality, as music consumption has undergone significant changes recently, especially since users are today just one click away from millions of songs [4]. This context results in the existence of large repositories of unstructured knowledge, which have great potential for musicological studies or tasks within MIR such as music recommendation.

In this paper, we put forward an integration procedure for enriching with music-related information a large dataset of Amazon customer reviews [18, 19], with semantic and acoustic metadata obtained from MusicBrainz [1] and AcousticBrainz [2], respectively. MusicBrainz (MB) is a large open music encyclopedia of music metadata, whist AcousticBrainz (AB) is a database of music and audio descriptors, computed from audio recordings via state-of-the-art Music Information Retrieval algorithms [26]. In addition, we further extend the *semantics* of the textual content from two standpoints. First, we apply an aspect-based sentiment analysis framework [7] which provides specific sentiment scores for different aspects present in the text, e.g. album cover, guitar, voice or lyrics. Second, we perform Entity Linking (EL), so that mentions to named entities such as Artist Names or Record Labels are linked to their corresponding Wikipedia entry [24].

This enriched dataset, henceforth referred to as Multimodal Album Reviews Dataset (MARD), includes affective, semantic, acoustic and metadata features. We benefit from this multidimensional information to carry out two experiments. First, we explore the contribution of such features to the Music Genre classification task, consisting in, given a song or album review, predict the genre it belongs to. Second, we use the substantial amount of information at our disposal for performing a diachronic analysis of music criticism. Specifically, we combine the metadata retrieved for each review with their associated sentiment information, and generate visualizations to help us investigate any potential trends in diachronic music appreciation and criticism. Based on this evidence, and since music evokes emotions through mechanisms that are not unique to music [16], we may go as far as using musical information as means for a better understanding of global affairs. Previous studies argue that national confidence may be expressed in any form of art, including music [20], and in fact, there is strong evidence suggesting that our emotional reactions to music have important and far-reaching implications for our beliefs, goals and actions, as members of social and cultural groups [1]. Our analysis hints at a potential correlation between the language used in music reviews and major geopolitical events or economic fluctuations. Finally, we argue that applying sentiment analysis to music corpora may be useful for diachronic musicological studies.

© Sergio Oramas[1], Luis Espinosa-Anke[2], Aonghus Lawlor[3], Xavier Serra[1], Horacio Saggion[2]. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Sergio Oramas[1], Luis Espinosa-Anke[2], Aonghus Lawlor[3], Xavier Serra[1], Horacio Saggion[2]. "Exploring Customer Reviews for Music Genre Classification and Evolutionary Studies", 17th International Society for Music Information Retrieval Conference, 2016.

[1] http://musicbrainz.org/
[2] http://acousticbrainz.org

2. RELATED WORK

One of the earliest attempts on review genre classification is described in [15], where experiments on multiclass genre classification and star rating prediction are described. Similarly, [14] extend these experiments with a novel approach for predicting usages of music via agglomerative clustering, and conclude that bigram features are more informative than unigram features. Moroever, part-of-speech (POS) tags along pattern mining techniques are applied in [8] to extract descriptive patterns for distinguishing negative from positive reviews. Additional textual evidence is leveraged in [5], who consider lyrics as well as texts referring to the meaning of the song, and used for training a kNN classifier for predicting song subjects (e.g. war, sex or drugs).

In [23], a dataset of music reviews is used for album rating prediction by exploiting features derived from sentiment analysis. First, music-related topics are extracted (e.g. artist or music work), and this topic information is further used as features for classification. One of the most thorough works on music reviews is described in [28]. It applies Natural Language Processing (NLP) techniques such as named entity recognition, text segmentation and sentiment analysis to music reviews for generating texts explaining good aspects of songs in recommender systems. In the line of review generation, [9] combine text analysis with acoustic descriptors in order to generate new reviews from the audio signal. Finally, semantic music information is used in [29] to improve topic-wise classification (album, artist, melody, lyrics, etc.) of music reviews using Support Vector Machines. This last approach differs from ours in that it enriches feature vectors by taking advantage of *ad-hoc* music dictionaries, while in our case we take advantage of Semantic Web resources.

As for sentiment classification of text, there is abundant literature on the matter [21], including opinions, reviews and blog posts classification as positive or negative. However, the impact of emotions has received considerably less attention in genre-wise text classification. We aim at bridging this gap by exploring aspect-level sentiment analysis features.

Finally, concerning studies on the evolution of music genres, these have traditionally focused on variation in audio descriptors, e.g. [17], where acoustic descriptors of 17,000 recordings between 1960 and 2010 are analyzed. Descriptors are discretized and redefined as descriptive words derived from several lexicons, which are subsequently used for topic modeling. In addition, [12] analyze expressions located near the keyword *jazz* in newswire collections from the 20th century in order to study the advent and reception of jazz in American popular culture. This work has resemblances to ours in that we also explore how textual evidence can be leveraged, with a particular focus on sentiment analysis, for performing descriptive analyses of music criticism.

3. MULTIMODAL ALBUM REVIEWS DATASET

MARD contains texts and accompanying metadata originally obtained from a much larger dataset of Amazon customer reviews [18, 19]. The original dataset provides millions of review texts together with additional information such as overall rating (between 0 to 5), date of publication, or creator id. Each review is associated to a product and, for each product, additional metadata is also provided, namely Amazon product id, list of similar products, price, sell rank and genre categories. From this initial dataset, we selected the subset of products categorized as *CDs & Vinyls*, which also fulfill the following criteria. First, considering that the Amazon taxonomy of music genres contains 27 labels in the first hierarchy level, and about 500 in total, we obtain a music-relevant subset and select 16 of the 27 which really define a music style and discard for instance region categories (e.g. World Music) and other categories non specifically related to a music style (e.g. Soundtrack, Miscellaneous, Special Interest), function-oriented categories (Karaoke, Holiday & Wedding) or categories whose albums might also be found under other categories (e.g. Opera & Classical Vocal, Broadway & Vocalists). We compiled albums belonging only to one of the 16 selected categories, i.e. no multiclass. Note that the original dataset contains not only reviews about CDs and Vinyls, but also about music DVDs and VHSs. Since these are not strictly speaking music audio products, we filter out those products also classified as "Movies & TV". Finally, since products classified as Classical and Pop are substantially more frequent in the original dataset, we compensate this unbalance by limiting the number of albums of any genre to 10,000. After this preprocessing, MARD amounts to a total of 65,566 albums and 263,525 customer reviews. A breakdown of the number of albums per genre is provided in Table 1.

Genre	Amazon	MusicBrainz	AcousticBrainz
Alternative Rock	2,674	1,696	564
Reggae	509	260	79
Classical	10,000	2,197	587
R&B	2,114	2,950	982
Country	2,771	1,032	424
Jazz	6,890	2,990	863
Metal	1,785	1,294	500
Pop	10,000	4,422	1701
New Age	2,656	638	155
Dance & Electronic	5,106	899	367
Rap & Hip-Hop	1,679	768	207
Latin Music	7,924	3,237	425
Rock	7,315	4,100	1482
Gospel	900	274	33
Blues	1,158	448	135
Folk	2,085	848	179
Total	66,566	28,053	8,683

Table 1: Number of albums by genre with information from the different sources in MARD

Having performed genre filtering, we enrich MARD by extracting artist names and record labels from the Amazon product page. We pivot over this information to query the MB search API to gather additional metadata such as release id, first release date, song titles and song ids. Mapping with MB is performed using the same methodology described in [25], following a pair-wise entity resolution

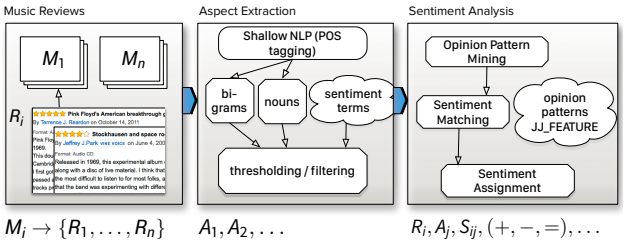

Figure 1: Overview of the opinion mining and sentiment analysis framework.

approach based on string similarity with a threshold value of $\theta = 0.85$. We successfully mapped 28,053 albums to MB. Then, we retrieved songs' audio descriptors from AB. From the 28,053 albums mapped to MB, a total of 8,683 albums are further linked to their corresponding AB entry, which encompasses 65,786 songs. The final dataset is freely available for download [3] .

4. TEXT PROCESSING

In this section we describe how we extract linguistic, sentimental and semantic information from textual reviews. This information will serve both as input features for our genre classification experiments, and also constitute the basis for the diachronic study described in Section 6.

4.1 Sentiment Analysis

Following the work of [6, 7] we use a combination of shallow NLP, opinion mining, and sentiment analysis to extract opinionated features from reviews. For reviews R_i of each album, we mine bi-grams and single-noun aspects (or review features), see [13]; e.g. bi-grams which conform to a noun followed by a noun (e.g. *chorus arrangement*) or an adjective followed by a noun (e.g. *original sound*) are considered, excluding bi-grams whose adjective is a sentiment word (e.g. *excellent*, *terrible*). Separately, single-noun aspects are validated by eliminating nouns that are rarely associated with sentiment words in reviews, since such nouns are unlikely to refer to item aspects. We refer to each of these extracted aspects A_j as review aspects.

For a review aspect A_j we determine if there are any sentiment words in the sentence containing A_j. If not, A_j is marked neutral, otherwise we identify the sentiment word w_{min} with the minimum word-distance to A_j. Next we determine the POS tags for w_{min}, A_i and any words that occur between w_{min} and A_i. We assign a sentiment score between -1 and 1 to A_j based on the sentiment of w_{min}, subject to whether the corresponding sentence contains any negation terms within 4 words of w_{min}. If there are no negation terms, then the sentiment assigned to A_j is that of the sentiment word in the sentiment lexicon; otherwise this sentiment is reversed. Our sentiment lexicon is derived from SentiWordNet [10] and is not specifically tuned for music reviews. An overview of the process is shown in Figure 1. The end result of sentiment analysis is that we determine a sentiment label S_{ij} for each aspect A_j in review R_i. A sample annotated review is shown in Figure 2

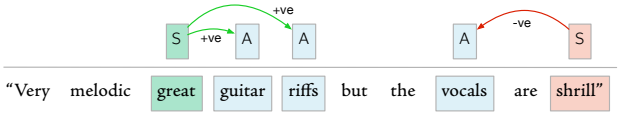

Figure 2: A sentence from a sample review annotated with opinion and aspect pairs.

4.2 Entity Linking

Entity Linking (EL) is the task to provide, given a mention to a named entity (e.g. person, location or organization), its most suitable entry in a reference Knowledge Base (KB) [22]. In our case, EL was performed taking advantage of Tagme [4] [11], an EL system that matches entity candidates with Wikipedia links, and then performs disambiguation exploiting both the in-link graph and the Wikipedia page dataset. TagMe provides for each detected entity, its Wikipedia page id and Wikipedia categories.

5. MUSIC GENRE CLASSIFICATION

5.1 Dataset Description

Starting from MARD, our purpose is to create a subset suitable for genre classification, including 100 albums per genre class. We enforce these albums to be authored by different artists, and that review texts and audio descriptors of their songs are available in MARD. Then, for every album, we selected audio descriptors of the first song of each album as representative sample of the album. From the original 16 genres, 3 of them did not have enough instances complying with these prerequisites (Reggae, Blues and Gospel). This results in a classification dataset composed of 1,300 albums, divided in 13 different genres, with around 1,000 characters of review per album.

5.2 Features

5.2.1 Textual Surface Features

We used a standard Vector Space Model representation of documents, where documents are represented as bag-of-words (BoW) after tokenizing and stopword removal. All words and bigrams (sequences of two words) are weighted according to *tf-idf* measure.

5.2.2 Semantic Features

We enriched the initial BoW vectors with semantic information thanks to the EL step. Specifically, for each named entity disambiguated with Tagme, its Wikipedia ID and its associated categories are added to the feature vector, also with *tf-idf* weighting. Wikipedia categories are organized in a taxonomy, so we enriched the vectors by adding one level more of broader categories to the ones provided by

[3] http://mtg.upf.edu/download/datasets/mard

[4] http://tagme.di.unipi.it/

	Alt. Rock	Classical	Country	Electronic	Folk	Jazz	Latin	Metal	New Age	Pop	R&B	Hip-Hop	Rock
Alt. Rock	28 / 42	1 / 3	3 / 1	10 / 10	7 / 1	1 / 2	2 / 0	18 / 12	10 / 2	4 / 10	3 / 6	3 / 2	10 / 9
Classical	0 / 0	87 / 95	1 / 0	0 / 0	1 / 1	1 / 1	2 / 2	1 / 0	5 / 1	1 / 0	0 / 0	0 / 0	1 / 0
Country	2 / 1	0 / 0	51 / 84	3 / 0	9 / 1	9 / 0	3 / 0	0 / 1	3 / 0	8 / 8	6 / 4	1 / 0	5 / 1
Electronic	7 / 3	3 / 1	1 / 2	40 / 61	4 / 1	1 / 2	2 / 2	6 / 0	7 / 5	6 / 5	6 / 7	13 / 5	4 / 7
Folk	4 / 6	11 / 0	13 / 10	7 / 0	27 / 55	6 / 1	7 / 3	4 / 2	6 / 9	5 / 9	6 / 4	1 / 0	3 / 1
Jazz	7 / 0	10 / 1	6 / 2	2 / 2	5 / 0	45 / 82	6 / 3	3 / 0	8 / 2	3 / 5	4 / 1	1 / 1	0 / 1
Latin	4 / 3	6 / 4	9 / 2	1 / 2	5 / 1	10 / 2	28 / 78	3 / 0	6 / 2	11 / 4	7 / 2	5 / 0	5 / 0
Metal	13 / 5	1 / 0	1 / 1	2 / 2	1 / 0	0 / 1	1 / 0	63 / 87	1 / 0	1 / 0	3 / 1	1 / 0	12 / 3
New Age	9 / 2	7 / 6	9 / 0	7 / 4	10 / 10	9 / 2	7 / 6	3 / 3	15 / 53	10 / 7	6 / 1	2 / 1	6 / 5
Pop	6 / 2	9 / 1	10 / 2	9 / 2	5 / 3	9 / 2	5 / 2	2 / 0	7 / 1	19 / 73	7 / 6	2 / 2	10 / 5
R&B	8 / 2	0 / 1	16 / 3	8 / 4	2 / 0	5 / 3	5 / 0	1 / 0	3 / 0	7 / 10	24 / 71	17 / 5	4 / 1
Hip-Hop	8 / 2	0 / 0	2 / 1	8 / 2	0 / 1	0 / 1	1 / 0	4 / 3	2 / 0	4 / 1	7 / 2	61 / 86	3 / 1
Rock	17 / 15	1 / 2	6 / 8	4 / 7	10 / 5	2 / 4	7 / 1	12 / 13	4 / 1	9 / 7	7 / 4	6 / 2	15 / 31

Table 2: Confusion matrix showing results derived from AB acoustic-based classifier/BoW+SEM text-based approach.

Tagme. Broader categories were obtained by querying DBpedia [5] .

5.2.3 *Sentiment Features*

Based on those aspects and associated polarity extracted with the opinion mining framework, with an average number of aspects per review around 37, we follow [21] and implement a set of sentiment features, namely:

- Positive to All Emotion Ratio: fraction of all sentimental features which are identified as positive (sentiment score greater than 0).
- Document Emotion Ratio: fraction of total words with sentiments attached. This feature captures the degree of affectivity of a document regardless of its polarity.
- Emotion Strength: This document-level feature is computed by averaging sentiment scores over all aspects in the document.
- F-Score [6] : This feature has proven useful for describing the contextuality/formality of language. It takes into consideration the presence of *a priori* "descriptive" POS tags (nouns and adjectives), as opposed to "action" ones such as verbs or adverbs.

5.2.4 *Acoustic Features*

Acoustic features are obtained from AB. They are computed using Essentia [7] . These encompass loudness, dynamics, spectral shape of the signal, as well as additional descriptors such as time-domain, rhythm, and tone [26].

5.3 Baseline approaches

Two baseline systems are implemented. First, we implement the text-based approach described in [15] for music review genre classification. In this work, a Naïve Bayes classifier is trained on a collection of 1,000 review texts, and after preprocessing (tokenisation and stemming), BoW features based on document frequencies are generated. The second baseline is computed using the AB framework for song classification [26]. Here, genre classification is computed using multi-class support vector machines (SVMs) with a one-vs.-one voting strategy. The classifier is trained with the set of low-level features present in AB.

[5] http://dbpedia.org

[6] Not to be confused with the evaluation metric.

[7] http://essentia.upf.edu/

	BoW	BoW+SEM	BoW+SENT
Linear SVM	**0.629**	**0.691**	**0.634**
Ridge Classifier	0.627	0.689	0.61
Random Forest	0.537	0.6	0.521

Table 3: Accuracy of the different classifiers

5.4 Experiments

We tested several classifiers typically used for text classification, namely Linear SVM, Ridge Classifier and Nearest Centroid, using the implementations provided by the scikit-learn library [8] . Among them, Linear SVM has shown better performance when combining different feature sets (see Table 3). Therefore, we trained a Linear SVM classifier with L2 penalty over different subsets of the features described in Section 5.2, which are combined via linear aggregation. Specifically, we combine the different feature sets into five systems, namely **BoW** (BoW), **BoW+Semantic** without broader categories (BoW+SEM), **BoW+Semantic Broader** with broader categories (BoW+SEMb), **BoW+Sentiment** (BoW+SENT) and **BoW+Semantic+Sentiment** (BoW+SEM+SENT). In this way, we aim at understanding the extent to which sentiment and semantic features (and their interaction) may contribute to the review genre classification task. Note that this paper is focused on the influence of textual features in genre classification, and classification based on acoustic features is simply used as a baseline for comparison. A proper combination of acoustic and textual features in text classification is a challenging problem and would require a deeper study that is out of the scope of this paper. The dataset is split 80-20% for training and testing, and accuracy values are obtained after 5-fold cross validation.

5.5 Results and Discussion

Accuracy results of the two baseline approaches introduced in Section 5.3 along with our approach variants are shown in Figure 3. At first sight, we may conclude that sentiment features contribute to slightly outperforming purely text-based approaches. This result implies that

[8] http://scikit-learn.org/

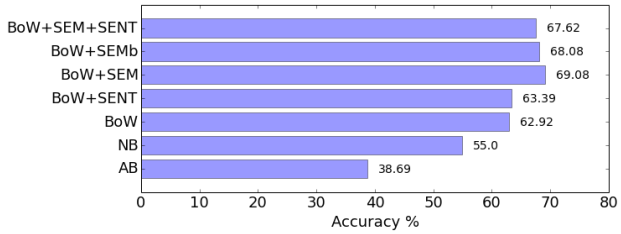

Figure 3: Percentage of accuracy of the different approaches. AB refers to the AcousticBrainz framework. NB refers to the method based on Naïve Bayes from [15].

affective language present in a music review is not a salient feature for genre classification (at least with the technology we applied), although it certainly helps. On the contrary, semantic features clearly boost pure text-based features, achieving 69.08% of accuracy. The inclusion of broader categories does not improve the results in the semantic approach. The combination of semantic and sentiment features improves the BoW approach, but the achieved accuracy is slightly lower than using semantic features only.

Let us review the results obtained with baseline systems. The Naïve Bayes approach from [15] is reported to achieve an accuracy of 78%, while in our results it is below 55%. The difference in accuracy may be due to the substantial difference in length of the review texts. In [15], review texts were at least 3,000 characters long, much larger that ours. Moreover, the addition of a distinction between Classic Rock and Alternative Rock is penalizing our results. As for the acoustic-based approach, although the obtained accuracy may seem low, it is in fact a good result for purely audio-based genre classification, given the high number of classes and the absence of artist bias in the dataset [3]. Finally, we refer to Table 2 to highlight the fact that the text-based approach clearly outperforms the acoustic-based classifier, although in general both show a similar behaviour across genres. Also, note the low accuracy for both Classic Rock and Alternative Rock, which suggests that their difference is subtle enough for making it a hard problem for automatic classification.

6. DIACHRONIC STUDY OF MUSIC CRITICISM

We carried out a study of the evolution of music criticism from two different temporal standpoints. Specifically, we consider when the review was written and, in addition, when the album was first published. Since we have sentiment information available for each review, we first computed an average sentiment score for each year of review publication (between 2000 and 2014). In this way, we may detect any significant fluctuation in the evolution of affective language during the 21st century. Then, we also calculated the average sentiment for each review by year of album publication. This information is obtained from MB and complemented with the average of the Amazon rating scores.

In what follows, we show visualizations for sentiment scores and correlation with ratings given by Amazon users, according to these two different temporal dimensions. Although arriving to musicological conclusions is out of the scope of this paper, we provide *food for thought* and present the readers with hypotheses that may explain some of the facts revealed by these data-driven trends.

6.1 Evolution by Review Publication Year

We applied sentiment and rating average calculations to the whole MARD dataset, grouping album reviews by year of publication of the review. Figure 4a shows the average of the sentiment scores associated to every aspect identified by the sentiment analysis framework in all the reviews published in a specific year, whilst Figure 4b shows average review ratings per year. At first sight, we do not observe any correlation between the trends illustrated in the figures. However, the sentiment curve (Figure 4a) shows a remarkable peak in 2008, a slightly lower one in 2013, and a low between 2003 and 2007, and also between 2009 and 2012. It is not trivial to give a proper explanation of this variations on the average sentiment. We speculate that these curve fluctuations may suggest some influence of economical or geopolitical circumstances in the language used in the reviews, such as the 2008 election of Barack Obama as president of the US. As stated by the political scientist Dominique Moïsi in [20]:

> In November 2008, at least for a time, hope prevailed over fear. The wall of racial prejudice fell as surely as the wall of oppression had fallen in Berlin twenty years earlier [...] Yet the emotional dimension of this election and the sense of pride it created in many Americans must not be underestimated.

Another factor that might be related to the positiveness in use of language is the economical situation. After several years of continuous economic growth, in 2007 a global economic crisis started [9], whose consequences were visible in the society after 2008 (see Figure 4c). In any case, further study of the different implied variables is necessary to reinforce any of these hypotheses.

.

6.2 Evolution by Album Publication Year

In this case, we study the evolution of the polarity of language by grouping reviews according to the album publication date. This date was gathered from MB, meaning that this study is conducted on the 42,1% of the MARD that was successfully mapped. We compared again the evolution of the average sentiment polarity (Figure 4d) with the evolution of the average rating (Figure 4e). Contrary to the results observed by review publication year, here we observe a strong correlation between ratings and sentiment polarity. To corroborate that, we computed first a smoothed version of the average graphs, by applying 1-D convolution (see line in red in Figures 4d and 4e). Then we computed Pearson's correlation between smoothed curves, obtaining a correlation $r = 0.75$, and a p-value $p \ll 0.001$. This means that in fact there is a strong correlation between

[9] https://research.stlouisfed.org

(a) Sentiment (b) Rating (c) USA GDP trend

(d) Sentiment (e) Rating (f) Sentiment by genre

Figure 4: Sentiment and rating averages by review publication year (a and b); GDP trend in USA from 2000 to 2014 (c), and sentiment and rating averages by album publication year (d, e and f)

the polarity identified by the sentiment analysis framework in the review texts, and the rating scores provided by the users. This correlation reinforces the conclusions that may be drawn from the sentiment analysis data.

To further dig into the utility of this polarity measure for studying genre evolution, we also computed the smoothed curve of the average sentiment by genre, and illustrate it with two idiosyncratic genres, namely *Pop* and *Reggae* (see Figure 4f). We observe in the case of *Reggae* that there is a time period where reviews have a substantial use of a more positive language between the second half of the 70s and the first half of the 80s, an epoch which is often called the golden age of *Reggae* [2]. This might be related to the publication of Bob Marley albums, one of the most influential artists in this genre, and the worldwide spread popularity of reggae music. In the case of *Pop*, we observe a more constant sentiment average. However, in the 60s and the beginning of 70s there are higher values, probably consequence by the release of albums by The Beatles. These results show that the use of sentiment analysis on music reviews over certain timelines may be useful to study genre evolution and identify influential events.

7. CONCLUSIONS AND FUTURE WORK

In this work we have presented MARD, a multimodal dataset of album customer reviews combining text, metadata and acoustic features gathered from Amazon, MB and AB respectively. Customer review texts are further enriched with named entity disambiguation along with polarity information derived from aspect-based sentiment analysis. Based on this information, a text-based genre classifier is trained using different combinations of features. A comparative evaluation of features suggests that a combination of bag-of-words and semantic information has higher discriminative power, outperforming competing systems in terms of accuracy. Our diachronic study of the sentiment polarity expressed in customer reviews explores two interesting ideas. First, the analysis of reviews classified by year of review publication suggests that geopolitical events or macro-economical circumstances may influence the way people speak about music. Second, an analysis of the reviews classified by year of album publication is presented. The results show how sentiment analysis can be very useful to study the evolution of music genres. The correlation observed between average rating and sentiment scores suggest the suitability of the proposed sentiment-based approach to predict user satisfaction with musical products. Moreover, according to the observed trend curves, we can state that we are now in one of the best periods of the recent history of music. Further work is necessary to elaborate on these hypotheses. In addition, the combination of audio and textual features is still an open problem, not only for classification but also for the study of the evolution of music. We expect the released dataset will be explored in multiple ways for the development of multimodal research approaches in MIR. In conclusion, the main contribution of this work is a demonstration of the utility of applying systematic linguistic processing on texts about music. Furthermore, we foresee our method to be of interest for musicologists, sociologists and humanities researchers in general.

8. ACKNOWLEDGEMENTS

This work was partially funded by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), by the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE), by the Keystone COST Action IC1302 and by the Insight Centre for Data Analytics under grant number SFI/12/RC/2289.

9. REFERENCES

[1] C S Alcorta, R Sosis, and D Finkel. Ritual harmony: Toward an evolutionary theory of music. *Behavioral*

and Brain Sciences, 31(5):576–+, 2008.

[2] Michael Randolph Alleyne and Sly Dunbar. *The Encyclopedia of Reggae: The Golden Age of Roots Reggae*. 2012.

[3] Dmitry Bogdanov, Alastair Porter, Perfecto Herrera, and Xavier Serra. Cross-collection evaluation for music classification tasks. In *ISMIR'16*, 2016.

[4] Òscar Celma and Perfecto Herrera. A new approach to evaluating novel recommendations. In *RecSys'08*, pages 179–186, 2008.

[5] Kahyun Choi, Jin Ha Lee, and J. Stephen Downie. What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 453–454, 2014.

[6] Ruihai Dong, Michael P O'Mahony, and Barry Smyth. Further Experiments in Opinionated Product Recommendation. In *ICCBR'14*, pages 110–124, Cork, Ireland, September 2014.

[7] Ruihai Dong, Markus Schaal, Michael P. O'Mahony, and Barry Smyth. Topic Extraction from Online Reviews for Classification and Recommendation. IJCAI'13, pages 1310–1316, 2013.

[8] J. Stephen Downie and Xiao Hu. Review mining for music digital libraries:phase II. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, page 196, 2006.

[9] Daniel P W Ellis. Automatic record reviews. *ICMIR*, 2004.

[10] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[11] Paolo Ferragina and Ugo Scaiella. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *Software, IEEE*, 29(1), June 2012.

[12] Maristella Johanna Feustle. Lexicon of Jazz invective: Hurling insults across a century with Big Data. *IAML/IMS'15*, 2015.

[13] Minqing Hu and Bing Liu. Mining Opinion Features in Customer Reviews. In *AAAI'04*, pages 755–760, San Jose, California, 2004.

[14] Xiao Hu and J Stephen Downie. Stylistics in customer reviews of cultural objects. *SIGIR Forum*, pages 49–51, 2006.

[15] Xiao Hu, J Stephen Downie, Kris West, and Andreas Ehmann. Mining Music Reviews: Promising Preliminary Results. *ISMIR*, 2005.

[16] Patrik N Juslin and Daniel Västfjäll. Emotional responses to music: the need to consider underlying mechanisms. *The Behavioral and brain sciences*, 31(5):559–621, 2008.

[17] Matthias Mauch, Robert M MacCallum, Mark Levy, and Armand M Leroi. The evolution of popular music: USA 1960-2010. *Royal Society Open Science*, 2015.

[18] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring Networks of Substitutable and Complementary Products. *KDD'15*, page 12, 2015.

[19] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based Recommendations on Styles and Substitutes. *SIGIR'15*, pages 1–11, 2015.

[20] Dominique Moisi. *The Geopolitics of Emotion: How Cultures of Fear, Humiliation, and Hope are Reshaping the World*. Anchor Books, New York, NY, USA, 2010.

[21] Calkin Suero Montero, Myriam Munezero, and Tuomo Kakkonen. *Computational Linguistics and Intelligent Text Processing*, pages 98–114, 2014.

[22] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *TACL'14*, 2:231–244, 2014.

[23] Tony Mullen and Nigel Collier. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. *EMNLP'04*, pages 412–418, 2004.

[24] Sergio Oramas, Luis Espinosa-anke, Mohamed Sordo, Horacio Saggion, and Xavier Serra. ELMD: An Automatically Generated Entity Linking Gold Standard in the Music Domain. *In LREC'16*, 2016.

[25] Sergio Oramas, Francisco Gómez, Emilia Gómez, and Joaquín Mora. Flabase: Towards the creation of a flamenco music knowledge base. In *ISMIR'15*, 2015.

[26] Alastair Porter, Dmitry Bogdanov, Robert Kaye, Roman Tsukanov, and Xavier Serra. Acousticbrainz: a community platform for gathering music information obtained from audio. *ISMIR'15*, pages 786–792, 2015.

[27] Maria Ruiz-Casado, Enrique Alfonseca, Manabu Okumura, and Pablo Castells. Information extraction and semantic annotation of wikipedia. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 145–169, 2008.

[28] Swati Tata and Barbara Di Eugenio. Generating Fine-Grained Reviews of Songs from Album Reviews. *Proceedings of the 48th ACL Anual Meeting*, (July):1376–1385, 2010.

[29] Wei Zheng, Chaokun Wang, Rui Li, Xiaoping Ou, and Weijun Chen. Music Review Classification Enhanced by Semantic Information. *Web Technologies and Applications*, 6612(60803016):5–16, 2011.